J Mol Model (2008) 14:441–449
DOI 10.1007/s00894-008-0293-0

ORIGINAL PAPER

Modeling the excitation wavelengths (λ_{ex}) of boronic acids

Minyong Li · Nanting Ni · Binghe Wang · Yanqing Zhang

Received: 3 December 2007 / Accepted: 18 February 2008 / Published online: 20 March 2008

Abstract The quantitative structure-property relationship (QSPR) method was used to model the fluorescence excitation wavelengths (λ_{ex}) of 42 boronic acid-based fluorescent biosensors (30 in the training set and 12 in the test set). In this QSPR study, unsupervised forward selection (UFS), stepwise multiple linear regression (SMLR), partial least squares regression (PLS) and associative neural networks (ASNN) were employed to simulate linear and nonlinear models. All models were validated by a test set and Tropsha's validation model. The resulting ASNN nonlinear model demonstrates significant improvement on the predictive ability of the neural network compared to the SMLR and PLS linear models. The descriptors used in the models are discussed in detail. These QSPR models are useful tools for the prediction of fluorescence excitation wavelengths of arylboronic acids.

Keywords QSPR · Boronic acids · Fluorescent biosensors · Excitation wavelength · UFS · SMLR · PLS · ASNN

M. Li (✉) · N. Ni · B. Wang (✉)
Department of Chemistry and Center
for Biotechnology and Drug Design, Georgia State University,
Atlanta, GA 30302–4098, USA
e-mail: mli@gsu.edu
e-mail: wang@gsu.edu

Y. Zhang
Department of Computer Sciences, Georgia State University,
Atlanta, GA 30302–3994, USA

Introduction

Since saccharides play very important roles in a wide range of biological processes [1, 2] and as biomarkers [3–5], a great deal of effort has been directed towards investigating ways to achieve selective recognition of various carbohydrates using small molecule probes/sensors [6–8]. In the search for small molecules probes/sensors for biologically important saccharides, the boronic acid group plays an especially important role as a key recognition moiety [9–12]. The ability of boronic acids to reversibly interact with diol-containing saccharides allows the boronic acid moiety to be used in carbohydrate recognition and sensing [13], and in saccharide and glycoprotein separation [14, 15]. In using the boronic acid moiety for carbohydrate sensor design it is especially important to have boronic acid reporter compounds that change their fluorescent properties upon binding [9, 16]. In designing such boronic acid fluorescent reporter compounds, the ability to estimate the excitation wavelength of an arylboronic acid is very important and can help save valuable research time. Therefore, we are interested in developing mathematical models for predicting the excitation wavelength of arylboronic acids.

So far the, mainly computational, efforts to predict fluorescence excitation wavelengths were based on quantum chemistry simulations, such as density function theory (DFT) [17, 18] and ab initio calculations [19, 20]. However, such accurate calculation of the fluorescence profiles is very time-consuming and complex, thus precluding the use of such methods to predict dozens of fluorescent compounds in a fast and accurate manner.

Quantitative structure-property relationship (QSPR) studies have been used successfully to predict the physico-

chemical properties of chemical compounds based on their structures [21, 22]. The biological counterpart of such studies, quantitative structure-activity relationship (QSAR), has also been extensively used with great success [23–25]. The concept of QSAR/QSPR is to transform searches for compounds with desired properties by chemical intuition and experience into a quantitative method using mathematical models [26]. Once a correlation model between a structure and an activity/property is found, any number of compounds, including those not yet synthesized, can be readily screened in silico to select structures with the desired properties. It is then possible to select the most promising compounds for synthesis and evaluation. Thus, QSAR/QSPR approaches conserve resources and greatly accelerate the process of developing new molecules for use as drugs, materials, additives, or for other proposes with high speed. Using this QSAR/QSPR approach, many efforts have successfully been made to investigate the spectral properties of various systems, for example the prediction of the excitation and emission maxima of green fluorescent protein (GFP) chromophores using an artificial neural network (ANN) [27], and the modeling of fluores-

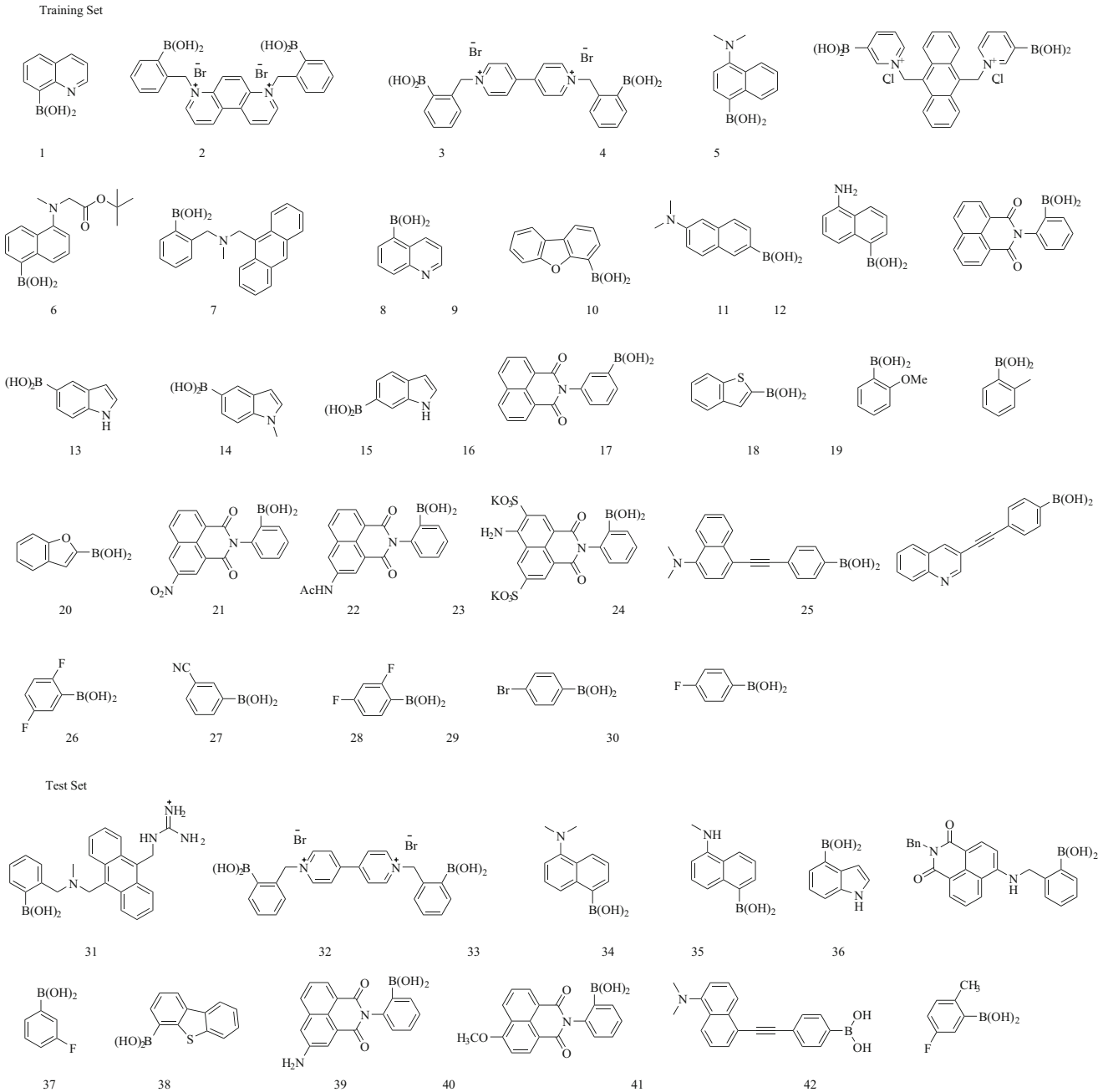

Fig. 1 Chemical structure of boronic acid-based biosensors

cence wavelengths of fluorescence probes using heuristic method (HM) and radial basis function neural networks (RBFNNs) [28].

In the present work, the QSPR approach was employed to predict the excitation wavelengths (λ_{ex}) of a diverse set of 42 boronic acid-based fluorescent biosensors. The E-DRAGON and MOPAC programs [29] were used for the generation of the various descriptors. Unsupervised forward selection (UFS) [30] was then utilized for the rational selection of descriptors. Finally, on the basis of the selected descriptors, stepwise multiple linear regression (SMLR) [31], partial least squares (PLS) regression [32] and associative neural network (ASNN) simulation [33] were performed for the development of quantitative linear and nonlinear QSPR models between excitation wavelengths (λ_{ex}) and chemical structures.

Materials and methods

Data set

The fluorescence excitation wavelengths of all 42 boronic acid-based biosensors were obtained from the literatures or by experiment. The structures of these boronic acids are given in Fig. 1. The experimental excitation wavelengths are listed in Table 1. The excitation wavelengths of all boronic acids were determined in phosphate buffer (pH 7.4) or aqueous solution. The data set is randomly divided into two sets: a training set of 30 compounds (1–30) and a test set of 12 compounds (31–42). The training set was used to select the descriptors and develop the QSPR models; the test set was then used to validate their accurate and predictive ability.

Fluorescence studies

Boronic acids were purchased from Frontier Scientific (Logan, UT) and Aldrich (Milwaukee, WI). Buffer reagents were obtained from Aldrich and Acros Organics (Fair Lawn NJ) and were used without purification. The water used for fluorescence studies was doubled distilled and further purified with a Milli-Q filtration system. Quartz cuvettes were used in all studies. A Shimadzu UV-1700 UV-visible spectrometer was used for all absorbance studies (Shimadzu, Kyoto, Japan). A Shimadzu RF-5301PC fluorometer was used for all fluorescent studies. Solutions of boronic acids (1×10^{-5} M) were prepared in 0.1 M phosphate buffer at pH 7.40. The solutions were then transferred to a 1 cm quartz cell and UV absorbances were recorded immediately. The fluorescence excitation wavelengths were measured as the UV maximum absorbance wavelengths.

Table 1 Descriptors and classes selected by unsupervised forward selection (UFS)

Descriptor	Class	Reference
PJI2[a]	Topological descriptors	[47]
nDB[b]	Constitutional descriptors	[48]
GATS1p[c]	2D autocorrelations	[49]
R7e+[d]	GETAWAY descriptors	[50]
BEHm1[e]	Burden eigenvalues	[51]
JGI4[f]	Topological charge indices	[52]
AROM[g]	Geometrical descriptors	[53]
MATS5m[h]	2D autocorrelations	[54]
MATS6m[i]	2D autocorrelations	[54]
MATS8m[j]	2D autocorrelations	[54]
MATS3p[k]	2D autocorrelations	[54]
DISPe[l]	Geometrical descriptors	[55]
RDF145m[m]	RDF descriptors	[56]
Mor13e[n]	3D-MoRSE descriptors	[57]
P1e[o]	WHIM descriptors	[58]
ALOGP[p]	Molecular properties	[59]
LUMO−HOMO[q]	Quantum descriptors	[29]

[a] 2D Petitjean shape index
[b] Number of double bonds
[c] Geary autocorrelation of path length 1 weighted by atomic polarizabilities
[d] GETAWAY R maximal autocorrelation of path length 7 weighted by atomic Sanderson electronegativities
[e] Highest eigenvalue n. 1 of Burden matrix weighted by atomic masses
[f] Mean topological charge index of order 4
[g] Aromaticity index
[h] Moran autocorrelation of path length 5 weighted by atomic masses
[i] Moran autocorrelation of path length 6 weighted by atomic masses
[j] Moran autocorrelation of path length 8 weighted by atomic masses
[k] Moran autocorrelation of path length 3 weighted by atomic polarizabilities
[l] d COMMA2 value weighted by atomic Sanderson electronegativities
[m] Radial Distribution Function of 14.5 weighted by atomic masses
[n] 3D-MoRSE of signal 13 weighted by atomic Sanderson electronegativities
[o] First component shape directional WHIM index weighted by atomic Sanderson electronegativities
[p] Ghose-Crippen octanol-water partition coefficient (logP)
[q] Energy gap between the lowest occupied molecular orbital energy (LUMO) and the highest occupied molecular orbital energy (HOMO)

Calculation of QSPR descriptors

The chemical structures of all compounds were drawn with the ChemDraw program (CambridgeSoft, Cambridge, MA). Geometry optimization was then performed with the semi-empirical quantum mechanics-based PM3 method [34–36] implemented the in MOPAC 7 program [29]. The MOPAC output files were used to calculate the descriptors using the E-DRAGON 1.0 online program (http://www.vcclab.org/lab/edragon/. All 1,664 descriptors generated by the E-DRAGON program could be described as belonging to one of 20 classes: constitutional descriptors, topological

Table 2 Descriptors, coefficients, standard error (SD) and *t*-values for the stepwise multiple linear regression (SMLR) linear model

Descriptor	Coefficient	SD
Constant	41.89	30.47
AROM	81.34	20.96
nDB	7.21	1.31
GATS1p	100.18	13.70
R7e+	-595.12	265.49
LUMO-HOMO	-19.82	1.53

descriptors, walk and path counts, connectivity indices, information indices, 2D autocorrelations, edge adjacency indices, Burden eigenvalue descriptors, topological charge indices, eigenvalue-based indices, Randic molecular profiles, geometrical descriptors, RDF descriptors, 3D-MoRSE descriptors, WHIM descriptors, GETAWAY descriptors, functional group counts, atom-centered fragments, charge descriptors and molecular properties. Meanwhile, nine quantum descriptors derived from the MOPAC calculation were also included in the QSPR construction [29]. The E-DRAGON descriptors were calculated on the online Virtual Computational Chemistry Laboratory (http://www.vcclab.org/) [37], and the MOPAC calculations were performed using MOPAC 7.1 program [38] on a Linux workstation.

Selection of descriptors

After identification of a large number of descriptors, rational selection was carried out to reduce the number of descriptors to an acceptable level containing no redundancies and minimal multicollinearity. In this selection, a novel descriptor reduction algorithm, unsupervised forward selection (UFS) [30], was employed to select suitable descriptors. UFS can select from a data matrix a maximal linearly independent set of columns with a minimal number of multiple correlations. It was designed for use in the development of QSAR models, where the m by n data matrix contains the values of n variables (typically molecular properties) for m objects (typically compounds). In descriptor selection, variables with small variance (not significant at the 5% level) were then removed. The UFS procedure was then applied repeatedly using values of R^2_{max} ranging from 0.1 to 0.9 in increments of 0.1, together with $R^2_{max} = 0.99$. In all cases, models were built from the subset of variables identified by UFS using the Portsmouth formulation of continuum regression (CR) [39], a procedure in which the model selection criterion depends on a continuous parameter α in the range $0 \leq \alpha \leq 1.5$. The CR calculations were performed with PARAGON software (can be downloaded from http://www.port.ac.uk) using values of R ranging from 0 to 1.5 in increments of 0.1. The UFS calculation was performed in the Virtual Computational Chemistry Laboratory (http://www.vcclab.org/) [37].

Stepwise multiple linear regression

As a commonly used statistical method in the QSAR/QSPR approach, stepwise multiple linear regression (SMLR) was employed in development of the linear QSPR model herein

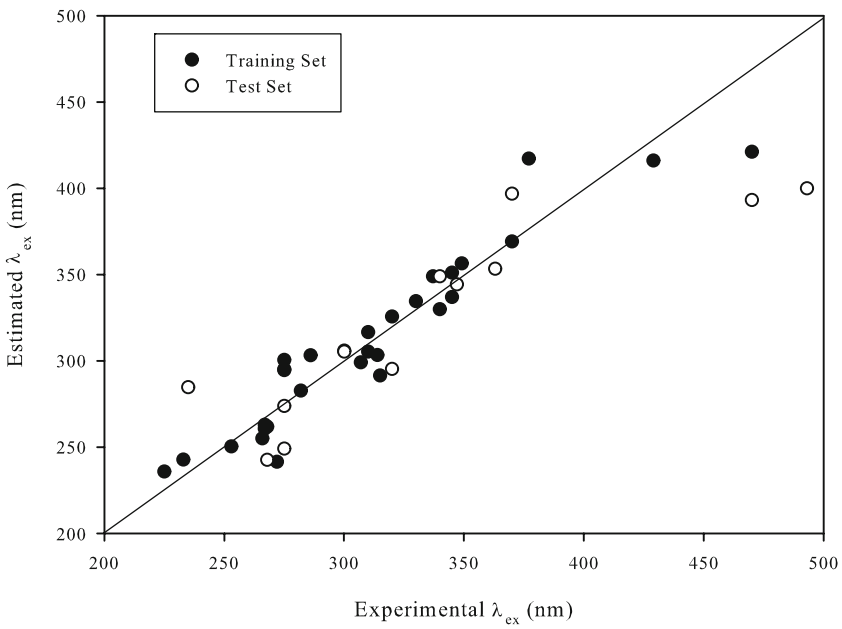

Fig. 2 Estimated vs experimental λ_{ex} by stepwise multiple linear regression (SMLR) linear model

Table 3 Descriptors, coefficients and SD for the partial least squares (PLS) linear model

Descriptor	Coefficient	SD
Constant	426.24	
nDB	3.051	0.92
GATS1p	26.62	24.24
BEHm1	1.25	6.61
JGI4	-1,045.89	352.12
AROM	70.99	130.05
MATS5m	32.74	5.04
MATS6m	39.77	23.71
R7e+	-565.05	241.35
LUMO−HOMO	-20.03	1.67

as implemented in the SAS 8.2 program [40] with all the default values. The statistical significance of stepwise addition of parameters was judged using the F value.

Partial least squares regression

It is well known that partial least squares (PLS) regression is quite sensitive to the noise created by excessive irrelevant descriptors in QSAR and QSPR modeling [32]. To achieve the best model quality, a two-step descriptor selection procedure was applied. The first step consists of the elimination of the low-variable (almost constant) descriptors that differ from a constant only for a few (2–3) compounds in the training set. Such descriptors cannot provide useful statistical information and simply help to fit these particular compounds, thus decreasing the predictivity. In the second step, the descriptor subset was optimized using Q^2-guided descriptor selection by means of a genetic algorithm. The stability and good prediction accuracy of this method has been demonstrated in computational experiments [41]. PLS analysis was calculated in the Virtual Computational Chemistry Laboratory (http://www.vcclab.org/) [37].

Associative neural network simulation

Associative neural network (ASNN) represents a combination of an ensemble of feed-forward neural networks and the *k*-nearest neighbor technique [33]. This method uses the correlation between ensemble responses as a measure of distance amongst the analyzed cases for the nearest neighbor technique, thus providing improved prediction through bias correction of the neural network ensemble. An associative neural network has a memory capacity in agreement with the training set. If new data becomes available, the network further improves its predictive ability and provides a reasonable approximation of the unknown function without the need to retrain the neural network ensemble. This feature dramatically improves its predictive ability over other traditional neural networks and *k*-nearest neighbor techniques. Another important feature of ASNN is the possibility to interpret neural network results by analyzing correlations between data cases in the space of

Table 4 Experimental and estimated excitation wavelengths (λ_{ex}). *ASNN* Associative neural network

No	Experimental	SMLR-estimated	PLS-estimated	ASNN-estimated	Reference
Training set					
1	314	303.44	292.59	298.45	[60]
2	470	421.21	421.55	461.92	[61]
3	300	305.89	314.32	301.55	[62]
4	377	417.14	403.23	384.31	[63]
5	320	325.68	335.65	323.17	[63]
6	370	369.21	376.47	374.01	*[a]
7	315	291.54	281.78	306.84	[64]
8	286	303.24	304.92	288.31	[65]
9	310	305.40	309.73	310.27	[65]
10	307	299.13	316.52	311.72	*
11	345	337.08	342.73	333.49	[66]
12	275	294.50	293.24	278.52	[67]
13	275	295.16	280.39	274.96	[67]
14	275	300.45	291.64	289.37	[67]
15	345	351.07	344.51	349.61	[66]
16	233	242.78	229.58	232.46	*
17	225	235.75	232.59	244.32	[68]
18	267	262.98	264.58	262.86	*
19	253	250.42	267.32	251.24	*
20	337	348.98	348.09	335.72	*
21	349	356.39	360.76	355.50	*
22	429	416.07	413.63	426.02	*
23	340	329.94	331.50	332.64	*
24	330	334.58	339.51	331.61	[69]
25	310	316.59	309.83	308.66	[69]
26	272	241.43	265.99	265.90	*
27	282	282.76	255.58	274.55	*
28	267	260.78	257.84	265.84	*
29	266	255.10	276.32	268.70	*
30	268	261.96	249.62	268.79	*
Test set					
31	370	396.83	403.12	388.12	[70]
32	470	393.15	414.56	474.62	[71]
33	300	305.30	321.66	293.22	[72]
34	320	295.32	326.82	304.44	*
35	275	273.86	321.30	288.93	*
36	493	399.94	392.20	415.46	[73]
37	268	242.61	249.75	259.69	*
38	235	284.70	290.14	262.11	*
39	347	344.21	353.86	354.05	*
40	363	353.36	352.74	358.24	*
41	340	348.98	353.45	298.79	[69]
42	275	249.26	244.13	258.74	*

[a] Values measured in our laboratory

models. In the current study, a standard type of neural network was employed for neural network training using the early stopping over ensemble (ESE) method. Training was stopped when a minimum error for the validation set was calculated ("early stopping" point). Neural network weights were updated using Levenberg-Marquardt algorithm. The number of hidden neurons was fixed at 3, and a 100 network ensemble was used. This ASNN simulation was performed using the program ASNN 1.0 [42] on a Windows workstation.

Quantitative structure-property relationship model validation

Typically, the final and most important characteristic of QSAR or QSPR model development is model validation, in which estimates of the predictive power of the model are generated [43]. Ideally, the predictive power should be defined as the ability of the model to accurately predict a target property, such as activity, of compounds that were not included in the training set in model development. Indeed, several recent publications suggest that the only way to ensure the high predictive power of a QSAR or QSPR model is to demonstrate a significant correlation between predicted and observed activities for a validation set of compounds that were not employed in model development [44, 45]. Therefore in our first validation, we used an external set containing 12 compounds to verify the accuracy of prediction.

Recently, Tropsha and coworkers have introduced a new set of validation criteria for QSAR or QSPR models [46, 47]. They consider a QSAR model predictive, if the following conditions are satisfied:

$$q^2 > 0.5$$
$$r^2 > 0.6 \quad \frac{(r^2 - r_0^2)}{r^2} < 0.1 \text{ or } \frac{(r^2 - r_0'^2)}{r^2} < 0.1$$
$$0.85 \le k \le 1.15 \text{ or } 0.85 \le k' \le 1.15$$

where r^2 is the correlation coefficient between the predicted and observed activities, r_0^2 the coefficient of determination between predicted and observed activities characterizing linear regression with the Y-intercept set to zero (i.e., described by Y = kX, where Y and X are the actual and predicted activity, respectively), $r_0'^2$ is the coefficient of determination between observed and predicted activities, and k and k' are the slopes of the regression lines through the origin.

It has been demonstrated that all of the above criteria are indeed necessary to adequately assess the predictive ability of a QSAR or QSPR model. Thus, we employed these criteria to validate the predictive power of our QSPR models.

Results and discussion

Unsupervised forward selection rational selection of descriptors

In the UFS phase, only 17 descriptors are significantly correlated with λ_{ex} at the 95% confidence interval (CI) level

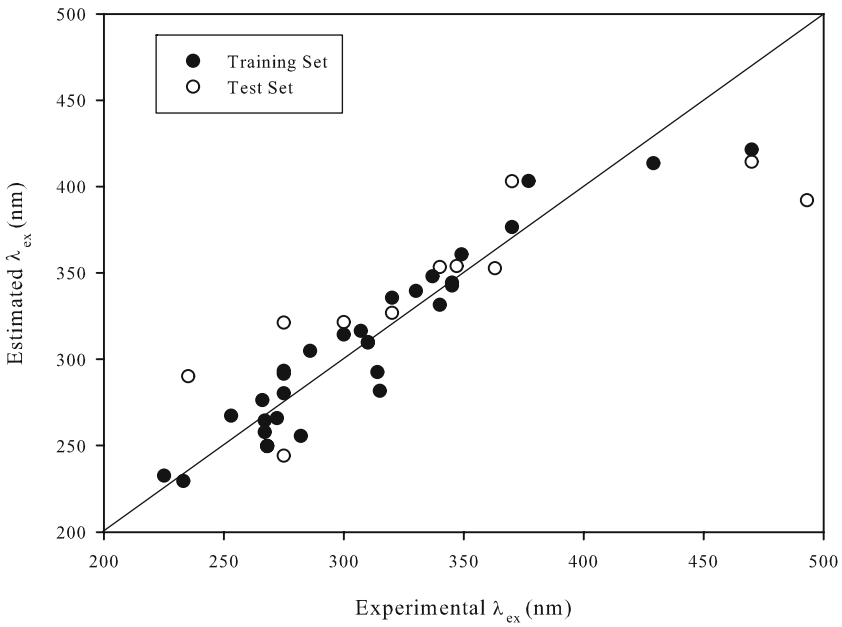

Fig. 3 Estimated vs experimental λ_{ex} by the partial least squares (PLS) linear model

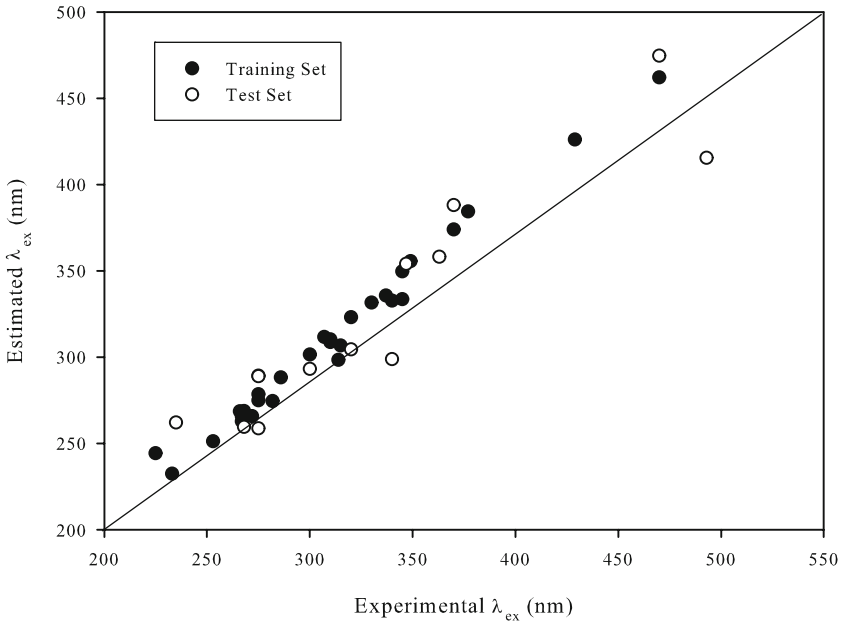

Fig. 4 Estimated versus experimental λ_{ex} by associative neural network (ASNN) nonlinear model

among the 1,664 descriptors used to model the excitation wavelength (λ_{ex}) QSPR. Belonging to 13 varied classes, these 17 descriptors were then used as input for the development of the linear and nonlinear QSPR models of λ_{ex}. The UFS-selected descriptors, classes and references are shown in Table 1.

Stepwise multiple linear regression linear model of λ_{ex}

Based on these 17 descriptors after selection, a training set of 30 compounds was used to develop an optimal SMLR linear model. Leave-one-out (LOO) cross-validation statistical parameters were calculated to evaluate the model quality. Finally, a five-descriptor (AROM, nDB, GATS1p, R7e+ and LUMO−HOMO) correlation model was obtained as represented in Table 2. The obtained squared correlation (r^2) was 0.899 and the LOO squared correlation (q^2) was 0.856 for the training set. The standard error (RMSE) was 17.45 and the *F*-value was 249.01. For the test set, the predicted results obtained were r^2 = 0.769, q^2 = 0.620, RMSE = 39.71 and *F*-value = 33.38. The estimated λ_{ex} values based on the SMLR linear model are listed in Table 4. Figure 2 depicts the estimated vs experimental λ_{ex} values for all 42 compounds.

Partial least squares linear model of λ_{ex}

A linear model of λ_{ex} was also developed by PLS using 17 selected descriptors. In this case, a correlation model consisting of nine descriptors (nDB, GASTS1p, BEHm1, JGI4, AROM, MATS5m, MATS6m, R7e + and LUMO−HOMO) was obtained as shown in Table 3. The number of PLS components is 2. For the training set, r^2 was 0.901 and q^2 was 0.880. RMSE was 17.31 and the *F*-value was 253.76. The predictive ability of the external test set yielded r^2 = 0.705, q^2 = 0.567, RMSE = 44.91 and *F*-value = 23.91. The estimated λ_{ex} results of the PLS model are shown in Table 4. The experimental and estimated λ_{ex} in the training and test set are shown in Fig. 3.

Associative neural network nonlinear model of λ_{ex}

In this case a nonlinear ASNN model of λ_{ex} was developed using the same selected subset of 17 descriptors as in the

Table 5 Validation parameters for three quantitative structure-property relationship (QSPR) models

Model	Training set		Test set		r_0^2	$(r^2 - r_0^2)$	*K*
	q^2	r^2	q^2	r^2		r^2	
SMLR	0.856	0.899	0.690	0.769	0.898	0.001	0.999
PLS	0.880	0.901	0.567	0.705	0.901	0	1.000
ASNN	0.983	0.983	0.858	0.874	0.982	0.001	1.001

linear models. The training of ASNN used Levenberg-Marquardt algorithm and k=10 nearest neighbors, and yielded an additional improvement in the results [r^2= 0.983, q^2=0.983, RMSE=7.10 and mean absolute error (MAE)=5.141] as compared with the linear models. Application of the ASNN model in the test set gave values of r^2=0.874, q^2=0.858, RMSE=29.67 and MAE=20.104. The estimated results of the ASNN nonlinear model are given in Table 4. Figure 4 shows the estimated vs experimental λ_{ex} using the ASNN nonlinear model. The ASNN has an RMSE of 7.10 nm for the training set and 29.67 nm for the test set. The r^2 of the training set is 0.983, and that of the test set is 0.874, whereas the value of q^2 is 0.983 and 0.858, respectively.

This ASNN model demonstrates a significant improvement in the predictive ability of a neural network compared to the SMLR and PLS linear models, thus indicating a nonlinear relationship between the descriptors and the fluorescence excitation wavelength (λ_{ex}).

Validation of QSPR models

In order to validate the QSPR models, Tropsha's validation criteria were applied as depicted in Table 5. It should be noted that all three models could be accepted as reliable QSPR models when judged by Tropsha's model [45, 46]. The best model—the ASNN nonlinear model—had q^2= 0.983 and r^2=0.983 for tthe raining set, and q^2=0.858 and r^2=0.874 for the test set, r_0^2 = 0.982 and k=1.001.

Interpretation of descriptors

To discuss the descriptors in the regression models, it is necessary to consider factors correlated with fluorescence wavelength λ_{ex}. In the case of the SMLR linear model, five descriptors—AROM, nDB, GATS1p, R7e+ and LUMO−HOMO—were included. Generally speaking, an increase in the extent of the π-electron system (i.e., degree of aromaticity and number of double bonds) leads to a shift in the absorption and fluorescence spectra to longer wavelengths [75]. The aromaticity index (AROM), as a geometric descriptor in all three nonlinear and linear models, reflects the planar and rigid geometry of the molecule, contributes positively to the excitation wavelength. The positive nature of this coefficient is in good accordance with the theory that the electron is easier to transfer in the molecule of aromatic plane. The number of double bonds (nDB) in both the two linear and the nonlinear models is a constitutional descriptor that measures the degree of unsaturation of the molecule. The positive coefficients of nDB are in good agreement with the fact that increasing the unsaturation of a molecule causes the chromophore to redshift. GATS1p, the Geary autocorrelation of path length 1 weighted by atomic polarizabilities, is related to conventional polarizability while allowing for attenuation of the influence of more remote atoms and bonds. The positive coefficient of GATS1p indicates that the fluorescence excitation wavelength would increase with increasing polarizability. Another affirmative descriptor is R7e+, the GETAWAY R maximal autocorrelation of path length 7 weighted by atomic Sanderson electronegativities. Atomic electronegativity reflects the electron-attracting ability of an atom in a particular molecular environment, and thus has a strong effect on fluorescence excitation wavelengths. The LUMO−HOMO energy gap is a quantum descriptor that approximates the energy difference between the electronic states. This descriptor relates directly to the fluorescence excitation wavelength, and thus appears in both linear models as well as the nonlinear model.

In the PLS linear model, four more descriptors, BEHm1, JGI4, MATS5m and MATS6m, were found to govern the fluorescence excitation wavelength. BEHm1, MATS5m and MATS6m are the descriptors weighted by atomic masses with positive contribution. The favorable nature of these coefficients suggests that the atomic mass may play an important role in the process of fluorescent excitation. As the mean topological charge index of order 4, JGI4 depicts the charge distribution. For this descriptor, the negative coefficient agrees well with R7e+ in the SMLR linear model.

Conclusions

The present report demonstrates that both linear and nonlinear QSPR models can be used successfully to predict fluorescence excitation wavelengths (λ_{ex}). All three models were validated by a test set and Tropsha's validation model [45, 46]. These models are easy to interpret and have high predictive ability. Among these three models, the ASNN nonlinear model demonstrates the most significant improvement in the predictive ability of a neural network compared to the SMLR and PLS linear models. The estimated results are in good agreement with experimental values. In conclusion, this QSPR approach can be used as a probe for the prediction of fluorescence excitation wavelengths, and the corresponding descriptors can also contribute to the fluorescent profiling of boronic acids.

Acknowledgments The authors thank Dr. Igor V. Tetko [University of Lausanne, Switzerland and Institute for Bioinformatics (MIPS)], who made available the ASNN 1.0 program. Some computations were done using Virtual Computational Chemistry Laboratory (http://www.vcclab.org). Financial support from the Georgia Cancer Coalition, Georgia Research Alliance, and the National Institutes of Health (CA123329, CA113917) is gratefully acknowledged.

References

1. Wells L, Vosseller K, Hart GW (2001) Science 291:2376–2378
2. Lehle L, Strahl S, Tanner W (2006) Angew Chem Int Ed Engl 45:6802–6818
3. Jeschke U, Mylonas I, Shabani N, Kunert-Keil C, Schindlbeck C, Gerber B, Friese K (2005) Anticancer Res 25:1615–1622
4. Wiest I, Schulze S, Kuhn C, Seliger C, Hausmann R, Betz P, Mayr D, Friese K, Jeschke U (2007) Anticancer Res 27:1981–1988
5. Pinho SS, Matos AJ, Lopes C, Marcos NT, Carvalheira J, Reis CA, Gartner F (2007) BMC Cancer 7:124
6. Wang W, Gao X, Wang B (2002) Curr Org Chem 6:1285–1317
7. Yan J, Fang H, Wang B (2005) Med Res Rev 25:490–520
8. Fang H, Kaur G, Wang B (2004) J Fluoresc 14:481–489
9. Yang W, Fan H, Gao X, Gao S, Karnati VV, Ni W, Hooks WB, Carson J, Weston B, Wang B (2004) Chem Biol 11:439–448
10. Striegler S, Dittel M (2003) J Am Chem Soc 125:11518–11524
11. Striegler S, Dittel M (2005) Inorg Chem 44:2728–2733
12. Shinkai S, Takeuchi M (2004) Biosens Bioelectron 20:1250–1259
13. Lin N, Yan J, Huang Z, Altier C, Li M, Carrasco N, Suyemoto M, Johnston L, Wang S, Wang Q, Fang H, Caton-Williams J, Wang B (2007) Nucleic Acids Res 35:1222–1229
14. Lee JH, Kim Y, Ha MY, Lee EK, Choo J (2005) J Am Soc Mass Spectrom 16:1456–1460
15. Monzo A, Bonn GK, Guttman A (2007) Anal Bioanal Chem
16. Ni W, Fang H, Springsteen G, Wang B (2004) J Org Chem 69:1999–2007
17. Somers KRF, Ceulemans A (2004) J Phys Chem A 108:7577–7583
18. Improta R, Santoro F (2005) J Phys Chem A 109:10058–10067
19. Hahn DK, Callis PR (1997) J Phys Chem A 101:2686–2691
20. Francisco JS, Li Y (2004) J Chem Phys 121:6298–6301
21. Jorgensen WL (2006) J Chem Inf Model 46:937
22. Dyekjaer JD, Jonsdottir SO (2004) Carbohydr Res 339:269–280
23. Selassie CD, Mekapati SB, Verma RP (2002) Curr Top Med Chem 2:1357–1379
24. Brown N, Lewis RA (2006) Curr Opin Drug Discov Devel 9:419–424
25. Kellogg GE, Semus SF (2003) EXS:223–241
26. Karelson M, Lobanov VS, Katritzky AR (1996) Chem Rev 96:1027–1044
27. Nantasenamat C, Isarankura-Na-Ayudhya C, Tansila N, Naenna T, Prachayasittikul V (2007) J Comput Chem 28:1275–1289
28. Shi J, Luan F, Zhang H, Liu M, Guo Q, Hu Z, Fan B (2006) QSAR Comb Sci 25:147–155
29. Stewart JJ (1990) J Comput Aided Mol Des 4:1–105
30. Whitley DC, Ford MG, Livingstone DJ (2000) J Chem Inf Comput Sci 40:1160–1168
31. Saxena AK, Prathipati P (2003) SAR QSAR Environ Res 14:433–445
32. Hasegawa K, Funatsu K (2000) SAR QSAR Environ Res 11:189–209
33. Tetko IV (2002) J Chem Inf Comput Sci 42:717–728
34. Stewart JJP (1989) J Comput Chem 10:209–220
35. Stewart JJP (1989) J Comput Chem 10:221–264
36. Stewart JJP (1991) J Comput Chem 12:320–341
37. Tetko IV, Gasteiger J, Todeschini R, Mauri A, Livingstone D, Ertl P, Palyulin VA, Radchenko EV, Zefirov NS, Makarenko AS, Tanchuk VY, Prokopenko VV (2005) J Comput Aided Mol Des 19:453–463
38. Stewart JJ (1993) MOPAC 7.1, Stewart Computational Chemistry, Colorado Springs, CO, http://openmopac.net/
39. Salt DW, Maccari L, Botta M, Ford MG (2004) J Comput Aided Mol Des 18:495–509
40. SAS 8.2, SAS Institute, Cary, NC
41. Palyulin VA, Radchenko EV, Zefirov NS (2000) J Chem Inf Comput Sci 40:659–667
42. Tekto IV, Kovalishyn VV (2001) ASNN 1.0, VCCLAB, Virtual Computational Chemistry Laboratory, http://www.vcclab.org
43. Kolossov E, Stanforth R (2007) SAR QSAR Environ Res 18:89–100
44. Golbraikh A, Tropsha A (2002) J Comput Aided Mol Des 16:357–369
45. Tropsha A, Gramatica P, Gombar VK (2003) QSAR Comb Sci 22:69–77
46. Golbraikh A, Tropsha A (2002) J Mol Graph Model 20:269–276
47. Golbraikh A, Shen M, Xiao Z, Xiao YD, Lee KH, Tropsha A (2003) J Comput Aided Mol Des 17:241–253
48. Petitjean M (1992) J Chem Inf Comput Sci 32:331–337
49. Todeschini R, Consonni V (2000) Handbook of molecular descriptors. Wiley-VCH, Berlin
50. Geary RC (1954) Incorp Statist 5:115–124
51. Consonni V, Todeschini R, Pavan M (2002) J Chem Inf Comput Sci 42:682–692
52. Burden FR (1989) J Chem Inf Comput Sci 29:225–227
53. Galvez J, Garcia R, Salabert MT, Soler R (1994) J Chem Inf Comput Sci 34:520–525
54. Kruszewski J, Krygowski TM (1972) Tetrahedron Lett 13:3839–3842
55. Moran PAP (1950) Biometrika 37:17–23
56. Silverman BD (2000) J Chem Inf Comput Sci 40:1470–1476
57. McCoy EF, Sykes MJ (2003) J Chem Inf Comput Sci 43:545–553
58. Schuur J, Gasteiger J (1997) Anal Chem 69:2398–2405
59. Bravi G, Gancia E, Mascagni P, Pegna M, Todeschini R, Zaliani A (1997) J Comput Aided Mol Des 11:79–92
60. Tetko IV, Tanchuk VY, Villa AE (2001) J Chem Inf Comput Sci 41:1407–1421
61. Yang W, Yan J, Springsteen G, Deeter S, Wang B (2003) Bioorg Med Chem Lett 13:1019–1022
62. Suri JT, Cordes DB, Cappuccio FE, Wessling RA, Singaram B (2003) Langmuir 19:5145–5152
63. Gao X, Zhang Y, Wang B (2003) Org Lett 5:4615–4618
64. Eggert H, Frederiksen J, Morin C, Norrild JC (1999) J Org Chem 64:3846–3852
65. James TD, Sandanayake KRAS, Iguchi R, Shinkai S (1995) J Am Chem Soc 117:8982–8987
66. Yang W, Lin L, Wang B (2005) Tetrahedron Lett 46:7981–7984
67. DiCesare N, Adhikari DP, Heynekamp JJ, Heagy MD, Lakowicz JR (2002) J Fluoresc 12:147–154
68. Wang J, Jin S, Lin N, Wang B (2006) Chem Biol Drug Des 67:137–144
69. Akay S, Yang W, Wang J, Lin L, Wang B (2007) Chem Biol Drug Des 70:279–289
70. Zheng S-L, Lin N, Reid S, Wang B (2007) Tetrahedron 63:5427–5436
71. Yang W, Yan J, Fang H, Wang B (2003) Chem Commun 792–793
72. Camara JN, Suri JT, Cappuccio FE, Wessling RA, Singaram B (2002) Tetrahedron Lett 43:1139–1141
73. Gao X, Zhang Y, Wang B (2005) New J Chem 29:579–586
74. Wang J, Jin S, Akay S, Wang B (2007) Eur J Org Chem 2091–2099
75. Valeur B (2000) Molecular fluorescence—an introduction: principles and applications. Wiley-VCH, Weinheim